Accepted Manuscript

From handcrafted to learned representations for human action recognition: a survey

Fan Zhu, Ling Shao, Jin Xie, Yi Fang

PII: S0262-8856(16)30108-1
DOI: doi: 10.1016/j.imavis.2016.06.007
Reference: IMAVIS 3525

To appear in: *Image and Vision Computing*

Received date: 15 September 2015
Revised date: 7 May 2016
Accepted date: 16 June 2016

Please cite this article as: Fan Zhu, Ling Shao, Jin Xie, Yi Fang, From handcrafted to learned representations for human action recognition: a survey, *Image and Vision Computing* (2016), doi: 10.1016/j.imavis.2016.06.007

From Handcrafted to Learned Representations for Human Action Recognition: A Survey

Fan Zhu[a,b], Ling Shao[c], Jin Xie[a,b], Yi Fang[a,b,*]

[a]*NYU Multimedia and Visual Computing Lab*
[b]*Electrical and Computer Engineering, New York University Abu Dhabi*
[c]*Department of Computer Science and Digital Technologies, Northumbria University*

Abstract

Human action recognition is an important branch among the studies of both human perception and computer vision systems. Along with the development of artificial intelligence, deep learning techniques have gained remarkable reputation when dealing with image categorization tasks (*e.g.,* object detection and classification). However, since human actions normally present in the form of sequential image frames, analyzing human action data requires significantly increased computational power than still images when deep learning techniques are employed. Such a challenge has been the bottleneck for the migration of learning-based image representation techniques to action sequences, so that the old fashioned handcrafted human action representations are still widely used for human action recognition tasks. On the other hand, since handcrafted representations are usually ad-hoc and overfit to specific data, they are incapable of being generalized to deal with various realistic scenarios. Consequently, resorting to deep learning action representations for human action recognition tasks is eventually a natural option. In this work, we provide a detailed overview of recent advancements in human action representations. As the first survey that covers both handcrafted and learning-based action representations, we explicitly discuss the superiorities and limitations of exiting techniques from both kinds. The ultimate goal of this survey is to provide comprehensive analysis and comparisons between learning-based and handcrafted action representations respectively, so as to inspire action recognition researchers towards the study of both kinds of rep-

[*] Corresponding author
Email address: yfang@nyu.edu (Yi Fang)

resentation techniques.

Keywords: Human action recognition, Handcrafted features, Deep Learning, Convolutional neural network, Dictionary learning

1. Introduction

Human perception is collectively the organization, identification and interpretation of feelings that humans acquire from the surrounding environment [1]. The investigation of human perception systems has been a challenging topic in many modern science subjects, including psychology [2], cognitive science [3], neuroscience [4] and biology [5]. At a low level, as one of human perceptions, the human vision system can receive a series of observations on an agent's body movements. Such observations are then passed to an intermediate level of human perception system, where predictions of movement categories can be made, e.g., running, waving hand, and jumping. Humans' daily activities are based on signals acquired from a large number of visual perceptions. Imagine that you are now playing basketball with another player on the play-ground. He is dribbling the ball in front of you while you are trying to defend him. At the lowest level of your visual perception system, most details of the other player's movements are perceived, e.g., his legs are straight, or his left hand is up and his right hand is down. At an intermediate level, you can tell that he is standing straightly and dribbling with his left hand. At the highest level, more advanced signals, such as he is more likely to pass the ball in the following move since his legs are not in a speeding up position, can be received based on intermediate perceptions. While one of the most important tasks of computer vision is to mimic the way how humans perceive the surrounding environment and make predictions accordingly, some existing computer vision technologies are biologically inspired, such as the Convolutional Neural Network (CNN) [6], which has gained a remarkable reputation for image-based categorization tasks in terms of performance. Similar as image classification, the ability of correctly recognizing human actions is a basic component of human perception system. In order to develop a robust action recognition system that can achieve a similar level of human performance, computer vision researchers have paid significant efforts in the past decades. Unfortunately, due to the challenging issues, such as high environment complexity and high intra-class action variations, what we can achieve today is still far from what a mature human perception system could do. Since the 1980s, when the research field of action recognition first drew the attention of computer science researchers, action representations mainly rely on statistics of gradients [7], combinations of global filters [8], depth images and skeletons [9], etc, which can be together referred as handcrafted features. Given the earlier stated importance of biologically inspired learning features, a growing number of learning-based video representations are altering the dominant position of old fashioned handcrafted features. In this survey, we aim to provide a comprehensive investigation of existing action representation and recognition approaches. In the remaining part of this section, we provide discussions on the scope of action recognition-related technologies that are covered in this survey and how the contents of this survey are structured.

1.1. Scope of study

Discussions in the survey concentrate on the study of action representation, which is seemingly a narrow investigation scope, but essentially is the core of action recognition. The term “action” is always confused with similar terms “gesture”, “interaction” and “activity”. To clarify, action is defined as intentional, purposive, conscious and subjectively meaningful activity, where the above stated four terms can be associated with a ascending order of complexity levels: “gesture”, “action”, “interaction” and “activity” [10]. Even though this survey mainly focuses on representation techniques which are based on action-related inputs, relevant representation methods (e.g., gesture representations), which are versatile for actions, are also covered. In the literature, there are several existing surveys on the topic of vision-based action recognition. These surveys are either structured following different taxonomies or emphasizing on different coverage areas. For example, both surveys by Aggarwal and Ryoo [10] and Cheng *et al.* [11] follow the taxonomy that divides action recognition approaches into single-layered approaches and hierarchical approaches; Moeslund *et al.* [12] and Poppe [13] follow the hierarchy of action primitive, action and activity; in an earlier work of Aggarwal and Cai [14], human motion analysis is discussed from three subtopics, which are 1) human body parts-based motion analysis, 2) moving human tracking and 3) image sequences-based human recognition. From another perspective, previous studies investigate the action recognition problem with different interest parts. For example, Cedras and Shah [15] mainly study motion-based action recognition approaches; Gavrila [16] investigate human body and hands tracking, and tracking-based recognition approaches. A more comprehensive summary of taxonomies and interests of studies based on previous relevant survey works are provided in Table 1. Due to the fact that human action is a collection of various human body movements, an obvious trend can be discovered from Table 1 that a significant number of work focuses on the investigation of human motions.

In recent years, the advent of large-scale training data has enabled CNN to significantly boost the performance of image classification on challenging tasks, such as ImageNet [17]. Such success has inspired researchers to follow a similar methodology to extract robust representations from action videos, so that an increasing number of methods that aim to utilize learning-based representations for action recognition have been proposed. While there exists a gap between the coverage of previous surveys on action recognition and the recently emerged learning-based action representations, our work only focuses on action representations and covers both handcrafted and learning-based approaches.

1.2. Survey structure

We structure this survey based on two main approaches of action representations, which are handcrafted approaches and learning-based approaches. In Section 2, we review the recent advancements of handcrafted action representations from four subcategories, including spatial-temporal volume-based approaches, depth image-based approaches, trajectory-based approaches and global approaches, where a discussion is given towards the end of Section 2 to summarize the cons and pros of listed handcrafted approaches. In Section 3, we provide a comprehensive investigation of existing learning-based action representation approaches, including both non-neural network-based approaches and neural network-based inspired approaches, where the focus is allocated to the latter. For non-neural network-based approaches, we selectively list genetic programming-based action representations and dictionary learning-based action representations, and for neural network-based approaches, a finer taxonomy is employed to review these approaches from the aspects of static frames-based approaches, frame transformations-based approaches, handcrafted features-based approaches, 3D CNN-based approaches and hybrid models, where certain overlapping may exist between these taxonomies. A discussion and comparison over both handcrafted action representations and learning-based action representations is also given towards the end of Section 3. Finally, a conclusion is given in Section 4.

Table 1: Summary of taxonomies and interests of studies based on previous relevant surveys.

Surveys	Taxonomy	Interest of study
Cheng *et al.* [11]	Single-layered approaches and hierarchical approaches	Handcrafted approaches
Moeslund *et al.* [12]	Hierarchy of action primitive, action and activity	Human motion capture and analysis
Aggarwal and Cai [14]	Human body parts-based motion analysis, moving human tracking and image sequences-based human recognition	Human motion capture and analysis
Aggarwal and Ryoo [10]	Single-layered approaches and hierarchical approaches	Human interaction and group activity analysis
Cedras and Shah [15]	Motion information extraction and motion-based recognition	Motion-based recognition approaches
Gavrila [16]	2D approaches with or without explicit shape models and 3D approaches	Human body and hands tracking-based motion analysis
Moeslund *et al.* [12]	Motion initialization, tracking, pose estimation and movement recognition	Human motion-based initialization, tracking, pose estimation and movement recognition
Poppe [13]	Hierarchy of action primitive, action and activity	Handcrafted action features and classification models
This work	Handcrafted and learning-based action representations	Action representations

2. Handcrafted action representations

Handcrafted action representations have been taking the dominated position along with the developments of action recognition. Before the emergence of deep learning approaches, the most popular action recognition frameworks follow an old-fashioned pipeline, which first extracts local statistics from both spatial saliences and action motions, then combines these local statistics into video-level representations and feeds to discriminative classifiers (such as Support Vector Machines (SVM) [18]). The low-level features are normally built on the pixel-level, and are carefully designed to deal with challenging issues, such as occlusions [19],[20] and viewpoint changes [21],[22],[23],[24],[25]. In the remaining part of this section, we review and summarize some popular handcrafted action features from the following aspects: 1) spatial-temporal volume-based approaches, 2) skeleton-based approaches, 3) trajectory-based approaches and 4) global approaches.

2.1. Spatial-temporal volume-based approaches

In those early ages of the time when action recognition problems start attracting computer vision researchers' attentions, the most popular way of developing an action recognition follows the common trend in object recognition that establishes the categorization frameworks based on detected spatio-temporal interest points [26]. Significant attempts are paid to develop effective detectors, including [26],[27],[28],[29]. While the majority of early interest point detectors are based on the extensions of scale invariant feature transformation (SIFT) [30]-like 2D interest point detectors, Dollar *et al.* [31] propose an alternative method of characterizing 3D cuboids of spatio-temporal volumes surrounding each detected 3D interest point, where interest points can be found by either detecting corners [32],[33] in the spatio-temporal space or using the Laplacian of Gaussian (LoG) [34] to obtain the local maximal response. Low-level action features, such as histograms of optical flows (HoF) [35], histograms of oriented gradients (HoG) [36], 3D extended HOG (HoG3D) [7], are then extracted from those sparsely detected spatial-temporal volumes and projected to video-level representations through coding schemes, such as the bag-of-words (BoW) model [37], sparse coding (SC) [38] or locality-constrained linear coding (LLC) [39].

While the sparse spatial-temporal volume approaches are technically sound and getting widely popularized, some later works suggest that improved performance can be achieved by adopting densely sampled spatial-temporal vol-

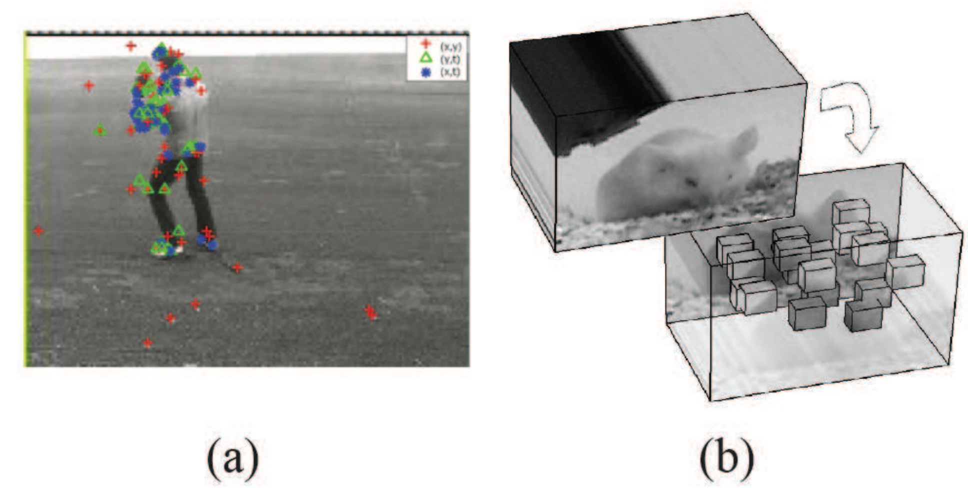

Figure 1: Illustration of (a): the spatial-temporal interest points-based approach, and (b): the spatial-temporal volume-based approach. (reprint from Gilbert *et al.* [29] and Dollar *et al.* [31], respectively).

umes. For example, Wang *et al.* [40] compare the performance of dense action representations and sparse representations in a local action feature evaluation work, where the BoW-based densely sampled approaches achieve better performance when either using HoG/HoF features or HoG3D features. Wainland *et al.* [19] also propose a HoG-based dense action representation that achieves outstanding performance when dealing with the occlusion problem in action recognition.

2.2. Depth image-based approaches

Due to the emergence and popularization of depth cameras [41],[42],[43] tasks that are relevant to pose estimation have been significantly simplified. A remarkable work that has to be mentioned is the real-time single image-based body part estimation algorithm [9], which is also the core technique that has been adopted by the human-machine interaction system of one of those best-seller gaming machines, Xbox, based on the Kinect depth camera [44]. Even that the work introduced in [9] is not directly applied to action recognition tasks, it has been the basis of the following depth input-based action recognition approaches. While previous depth image-based approaches suffer from difficulties in either the efficiency or the re-initialization precess,

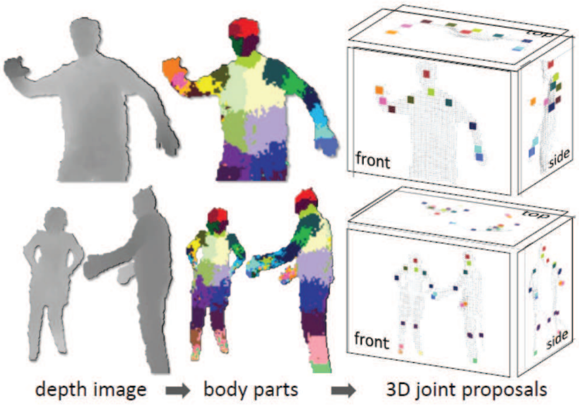

Figure 2: Illustration of how skeletons can be obtained based on depth images. (reprint from Shotton *et al.* [9]).

the proposed method can achieve fast and robust body part estimations based on single images and consumer hardwares. The body part estimation task is treated as a depth image pixel-level classification task, where the computational efficiency is mainly guaranteed by the utilization of random decision forests. Reliable body part proposal that describe positions of 3D skeleton joints can be obtained at the output end of the algorithm. An illustration of how human skeletons can be obtained based on depth images is given inf Figure. 2.

2.3. Trajectory-based approaches

Action trajectories can be computed by tracking human body joints along the action videos. Over the recent years, trajectory-based action representations have been proved as the most successful handcrafted shallow representations. Wang *et al.* [45] introduce a densely sampled rich feature that integrates HOG, HOF and motion boundary histogram (MBH) [46], where trajectories are obtained by tracking densely sampled points using optical flow. Since the global smoothness constraints are imposed to the trajectories, the resulting trajectories are much more robust. After the emergence of the dense trajectories-based action representation, a large amount of attempts have been paid, targeting on further improve the action recognition performance with more advanced trajectories-based approaches. Along with the improved performance when incorporating with a high level of density of trajectories within the video, high computational costs can be naturally anticipated in the end. In order to alleviate the computation complexity, Vig *et al.* [47] employ saliency-maps to extract salient regions within the frames where actions are performed. Based on the saliency-maps derived from human eye movements, a significant amount of dense trajectories can be discarded, while the action recognition performance can still remain at a high level. Improved performance over [45] can be even observed when pruning proportions that vary between 20% $\sim$ 50%. In order to enhance action representations against camera movements, Jiang *et al.* [48] utilize both local and global reference points that operate on the top of action trajectories to characterize action motions. Similarly, by investigating motions generated by camera movements, Wang *et al.* [49] further improve the performance of their previous dense trajectory work [45] by differentiating motions caused by human movements from motions caused by camera movements, where camera motions can be estimated by matching feature points across different frames using the speeded up robust features (SURF) [50] and dense optical flows [35]. Just like the dense trajectories [45], the improved dense trajectories [49] soon become the most popularized action representations among all handcrafted approaches. On the top of the improved dense trajectory features, Peng *et al.* [51] further build a multi-layer stacked fisher vector (FV) [52] approach that stacking FVs obtained from the improved dense trajectory features within sub-volumes to boost the performance.

2.4. Global approaches

Almost all the above mentioned approaches operate within a local video volume. On the contrary, there are also some global approaches that extract action representations from the holistic action videos. While local approaches can well preserve the local structures within video volumes and being insensitive to partial occlusions, the main advantage of global approaches, on the other hand, is to well capture the global structure of actions. Also, since global approaches normally take the whole action video as the framework input and perform pooling or filtering on the video pixel level, it always results in a relatively simple architecture and thus requires less computational costs. Zhen *et al.* [53] propose a spatio-temporal steerable pyramid (STSP) action representation. By decomposing spatio-temporal volumes into band-passed sub-volumes, spatio-temporal patterns can present in multiple pyramid scales. Three-dimensional separable steerable filters are then employed to the band-passed sub-volumes, and the outputs of these filters are squared, summed and max-pooled to generate video-level action representations.

Shao *et al.* [54] present a spatio-temporal Laplacian pyramid coding (STLPC) method for obtaining holistic action representations. STLPC decomposes each video sequence into a set of band-pass-filtered components and localizes spatio-temporal pyramid features that reside at different scales, so that the motion information can be effectively encoded. After the band-pass-filtering stage, a bank of 3D 3D-Gabor filters [55] is employed to each level of the Laplacian pyramid, where max pooling is employed to each filter band over spatio-temporal neighborhoods in order to capture both spatial structure and temporal motion information.

Some learning-based approaches [56],[57],[58] also share the characteristics as the above mentioned global approaches, e.g., inputting the whole video sequence and capturing the global structure, however, we exclude these learning-based approaches from this subsection and place them in Section 3.

2.5. Discussions

Along with the developments of handcrafted action features, previous approaches pay most efforts on capturing the motion information embedded in sequential body movements. The current successes of handcrafted features are achieved step-by-step by a wide range of techniques, such as optical flows, MBH, HOG, HOF and dense trajectories. So far, the most effective action feature is the IDT [49] and its extended stacking version [51] that works along with fisher vectors. The successes of handcrafted are also impacting learning-based action representations. Some existing learning-based action representation is built on handcrafted features. On the other hand, handcrafted action features also have some limitations. Since the most effective existing handcrafted features are local features, and follow the densely sample strategy (such as IDT), one major limitation for these approaches is the high computation complexity that are unavoidable in both the training and testing phases. Such a limitation can hinder handcrafted action features from many real-time applications.

3. Learning-based action representations

While handcrafted action features normally operate at video pixel levels and measure the low-level statistics of either spatial body shapes or temporal motions, recently, more advanced approaches that either operate on the top of handcrafted action features or establish end-to-end action recognition frameworks from the pixel-level to action categories have been proposed. We summarize existing learning-based action representations as non-neural network-based representations and neural network-based representations. Intuitively, the difference with non-neural network-based approaches and neural network-based approaches is that the latter is designed to mimic the way of how humans observe the world from a biological perspective. As introduced in the Section 1, deep learning approaches, such as CNN, consist of multiple layers, where these layers are expected to function in a similar manner as neural layers of the human perception system.

3.1. Genetic programming-based approaches

Genetic programming (GP) [59], as an evolutionary method, has been employed to a wide range of visual categorization tasks. Following the Darwinian principle of natural selection to automatically search a space of possible solutions without any prior knowledge, GP relies on a natural and random process that escapes traps by which handcrafted approaches may be captured. In a GP algorithm, a group of primitive operators is first employed to randomly assemble computational programs as initialization. Then, evolutions are performed over the population using either crossover or mutation strategies through reproduction with single or pair parents chosen stochastically [60]. Liu *et al.* utilize a population of primitive 3D operators, such as 3D-Gabor filter [55] and wavelet [61], to evolve the motion features from both colors and optical flow fields. The GP fitness function that calculates the average cross-validation classification error using a SVM on the training data is employed in the evolutionary architecture. By finishing the entire evolution procedure, the selected solution is a set of cascaded operator that can directly perform on action sequences, which results in an efficient feature extraction step in the querying phase.

3.2. Dictionary learning-based approaches

Dictionary learning is a popular type of representation learning method, which in most cases refers to learning sparse representations of the input data in the form of a linear combination of basis dictionary atoms. Since researchers find that the sparse representation of signals are particularly effective when dealing with data categorization tasks, dictionary learning techniques have been widely applied in a wide range of computer vision applications, including image classification, saliency detection, action recognition.

The well known Bag-of-Words (BoW) model [62, 63] is based on a particular type of dictionary learning method, which, instead of using a combination of basis dictionary atoms, assigns each input signal to a single basis dictionary atom. The BoW model is widely used by local approaches to generate global data representations, thus the BoW model can be employed by the majority approaches introduced in Section 2.1, 2.2 and 2.3. In order to obtain more discriminative action representations through sparse coding, Guha *et al.* [64] propose to learn sparse representation for action recognition and show improved performance over the BoW model. The classical objective function for sparsity-based dictionary learning contains a reconstruction error term and a regularization term, where the later penalizes on the number of selected basis dictionary atoms using l_0-norm or l_1-norm. Beyond the classical sparsity-based dictionary learning approaches, some variants are proposed to deal with advanced requirements, such as enforcing discriminative power to the learned sparse codes and minimizing cross-domain discrepancies for transfer learning tasks. Zheng *et al.* [65] learn a cross-view dictionary pair over actions captured from different observation points, and encode actions within each specific view with the corresponding dictionary, so that the encoded action representations can be view-invariant. Zhu and Shao [66] present a weakly-supervised dictionary learning approach to adapt knowledge of one action dataset to the other action dataset. In addition to the commonly used reconstruction error term, the dictionary learning function has a discriminative term and a cross-domain discrepancy term, so that the cross-domain smoothness property can be guaranteed in the learned action codes.

3.3. Neural network-based approaches

The main focus of this work is neural network-based features, which will be explicitly discussed in the remaining part of this section. As a popular biologically inspired technique nowadays, deep learning is one machine learning algorithm that attempts to model high-level abstraction of data using hierarchical structures. In contrast to handcrafted features, deep learning performs more intellectual leaning and contains hierarchical feature extraction layers that contain much more trainable parameters than shallow architectures, e.g., kernel machines [67]. Based on the success of applying deep learning techniques to image-level categorization tasks, some recent works have been using similar learning-based representations for action recognition, where these works are summarized according to the following directions: 1) learning from video frames, 2) learning from frame transformations, 3) learning from handcrafted features, 4) three-dimensional convolutional networks and 5) hybrid models.

3.3.1. Learning from static frames

While there are several existing well established deep learning frameworks for image feature extraction, e.g., OverFeat [68] and Caffe [69], applying any of these frameworks to frames within an action video and extract action features at the frame-level can be a natural and the most straightforward option.

Ning *et al* [70] decompose the video-based analysis of embryos development problem into frame-level 2D images, and apply two-dimensional CNN to various stages (from fertilization to four-cell stage) of the development process.

3.3.2. Learning from frame transformations

A Restricted Boltzmann Machine (RBM) [71] is a generative stochastic artificial neural network that can provide a “deep architecture” by successively composing several RBMs. Based on RBM, Taylor *et al.* [72] propose an unsupervised approach for learning spatio-temporal features (STF) using a gated Restricted Boltzmann Machine (GRBM)-based [73] convolutional architecture [6] to extract motion-sensitive action features from neighboring image pairs. Specifically, the GRBM architecture is first proposed by Memisevic and Hinton [6] to describe the probabilistic model of learning rich and distributed representations of image transformations. Such a generative model tries to predict the next frame image in a stream of observations based on the current frame, so as to perform “mapping” between both frames, and subsequently extract stream features based on frame transformations. By extending the GRBM model to image patches at identical spatial locations in sequential frames, GRBM can naturally capture the transformations of successive image pairs. In order to avoid the limitations of training GRBMs on isolated patches as in [6] and [74], Taylor *et al.* [72] extend a GRBM to a convolutional GRBM (convGRBM) by incorporating with the convolutional architecture [6],[75], where weights at multiple locations within an image are shared by a convGRBM. The convGRBM operates in a multi-stage architecture. At the lowest layer, convGRBM extracts motion-sensitive features from every neighboring frame pairs. At the intermediate layer, spatio-temporal cues are captured by 3D spatio-temporal filters, which are extended from the 2D spatial filter by Jarrett *et al.* [76]. After performing normalization, average spatial pooling and an additional max pooling in the temporal dimension, action representations can be obtained by the fully-connected layers and action label is at the topmost layer (softmax). The low layer conGRBM is trained unsupervised and separately with upper layers, where backpropagations are performed on upper layers.

3.3.3. Learning from handcrafted features

Another natural choice of computing learning-based action representations is to perform learning on the top of handcrafted features. In fact, the majority of early attempts that aim to address action recognition problem using learning-based representations lie in this category. This subsection needs to be distinguished from the previous section, "handcrafted representations-based action recognition", because the former discusses how the learning network can be established based on existing low-level features while the latter discusses how low-level features can be extracted from raw pixel-level video data. Kim *et al.* [77] propose a modified convolutional neural network (MCNN)-based action feature extraction and classification framework. At the low-level, action information are captured by handcrafted features. Specifically, When an action performed by an agent is presented in a 3D video, a sequence of the agent's outer boundary, which can be considered as a 2D contour in the spatial plane, generates a spatio-temporal volume, where the outer boundary information are extracted using three-dimensional Gabor filters [78]. Thus, in each spatio-temporal volume, actions are presented in a view-invariant form. In order to diminish the post-normalization location variances, the extended three-dimensional convolutional neural network is applied to each spatio-temporal volume, and subsequently action features can be extracted from a set of hierarchical layers based on the agent's outer boundary features. A 3DCNN consists two convolution layers and two subsampling layers, where each convolution layer or subsampling layer consists of two sub-layers. The extracted features are then fed into a discriminative classification model [79].

Jhuang *et al.* [80] present a biologically inspired system (BIS) that utilizes a feedforward hierarchy of spatio-temporal feature detectors of increasing complexity to measure motion-direction sensitive units, which lead to position-invariant spatio-temporal feature detectors. Motivated by the scale and position invariant features [81], [6], [82], [83], a vector of scale and position invariant features is obtained by computing a global max for each feature map at the top of the hierarchy. The main weakness of this approach is that it requires carefully handcrafted spatio-temporal feature detectors.

Wu and Shao [84] propose a hierarchical parametric networks (HPN) based on skeleton features. By replacing the RBM in [85] with a multilayer network, the HPN approach can serve as a better model for estimating emission probability of hidden Moarkov models [86] and achieve improved performance over other well established methods. Similar as [9], the approach introduced in [84] also operates on depth images.

3.3.4. Three-dimensional convolutional networks

The first attempt that aims to develop a three-dimensional convolutional neural networks (3D CNN) and performs 3D convolution along both spatial and temporal dimensions at the pixel level is introduced in [56]. By applying multiple distinct convolutional operations at identical input locations, and subsequently extracting features from multiple information channels, action representations obtained by such a 3D CNN approach contain a variety of information. In a 2D CNN [6], convolution is performed in the spatial domain, where features are extracted from neighboring units that share the same feature map in the previous layer. On the other hand, convolution is performed in both spatial and temporal dimensions using 3D cubes, which are generated by stacking multiple contiguous frames. The proposed architecture of 3D CNN consists of 7 layers including the input layer, which is hardwired to three convolution layers and two subsampling layers in with an alternating order. The last layer consists of 128 feature maps, and is fully connected to all feature maps in the previous layer. The main premise of the 3D CNN architecture comes from the feed-forward nature, which enables efficient feature extraction in the recognition phase. An illustration of the 3D CNN architecture used in [56] is given in Figure. 3.

Among the first batch of attempts that address action representations with learning-based methods, Le *et al.* [57] propose the hierarchical invariant spatio-temporal (HIST) action features. The feature learning framework is established based on the independent subspace analysis (ISA) [87], which is a two-layered network and normally used for extracting features from 2D images. Since the ISA training process is less efficient, especially when handling large-scale video data, the migration of ISA to the video domain is assisted by the principle component analysis (PCA) [88]. The complete HIST framework consists of several ISAs, where each ISA is firstly trained on small input patches (flattened vectors from sequences of image patches) and then propagate responses to the next-layer ISA with reduced dimensions by PCA.

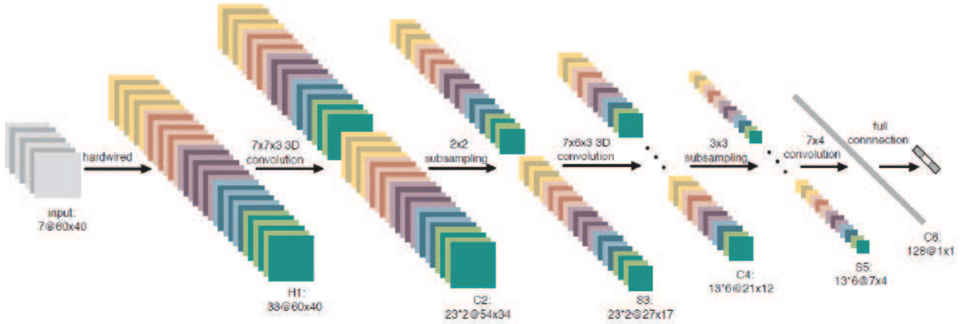

Figure 3: The 3D CNN architecture used in [56]. (reprint from Ji *et al.* [56]).

Inherit from the unsupervised learning property of ISA, HIST also operates in an unsupervised manner over action videos, so that it can leverage the massively available unlabeled online video data.

Baccouche *et al.* [58] present sequential deep learning (SDL) approach that adopts a similar strategy of extending CNN to a three-dimensional scenario as in Kim *et al.* [77]. Such an extension is rather straightforward in [58], since the two-dimensional convolutions are simply replaced by three-dimensional convolutions. On the other hand, [58] differs from [77] in a way that the CNN architecture directly operates on raw video pixels in the former scenario while the CNN framework is established on the top of handcrafted low-level features in the latter. The construction of the CNN architecture in [58] is also different from those in [77] by following the order of two alternating convolutional layers, a rectification layer, a sub-sampling layer, a second alternating convolutional layer, a second sub-sampling layer, a third convolutional layer and two neuron layers. The training process follows a standard online backpropagation with momentum algorithm [6]. Once action features are extracted using the three-dimensional CNN architecture, a sequential action labeling scheme is utilized. Instead of utilizing small sized spatio-temporal volumes to generate three-dimensional regions for CNN learning, Baccouche *et al.* [58] capture the features' temporal evolution over time by adapting CNN to sequential data, where the recurrent neural networks (RNN) [89] with one hidden layers of Long Short-Term Memory (LSTM) [90]. Apart from the LSTM model [58], some other RNN-based approaches are proposed for human action recognition tasks, including the hierarchical recurrent neural network for skeleton-based representation [91] and the differential recurrent neural networks [91]. However, when comparing with the LSTM model, regular RNN is incapable of capturing long-term dependencies between frames, so that these models are theoretically inferior.

Karpathy *et al.* [92] conduct a comprehensive study that provides three CNN-based action recognition approaches and extensively evaluates all these approaches. The main focus of this study is to investigate the best approach to incorporate with the motion information for recognizing actions, and how much improvement the motion information can boost from applying CNNs on static video frames. Based on extensive investigations on a new large-scale action video dataset, which consists of 1 million YouTube videos belonging to 487 classes, and many other real-world on-line videos, Karpathy *et al.* [92] discover that most videos normaly present in a highly inconsistent nature, and thus cannot be easily processed with fixed-sized architectures. Consequently, the network is designed to learn spatio-temporal features by connecting several contiguous frames in time and plugging in the network. In order to better analyze the benefits that come from the motion information, Karpathy *et al.* [92] present three CNN-based learning strategies: Early Fusion, Late Fusion and Slow Fusion. Illustrations of these fusion strategies are given in the reprinted Figure. 4 from [92]. Based on the single frame architecture, which is equivalent to applying CNNs to 2D images, Late Fusion enforces two separate CNNs on two apart frames (e.g., from a distance of 15 frames) and connects both networks in the first fully connected layer, so that action motion characteristics can be captured at a global level. Early Fusion, on the other hand, modifies the 2D single-frame based convolution window to include the temporal dimension, and feed these 3D cubes to the first convolutional layer. The Early Fusion strategy is equivalent to the work [6] and [58], in terms of how the motion information are captured. As a compromise between Early Fusion and Late Fusion, Slow Fusion (SFCNN) progressively connects adjacent frames from convolutional layers in both spatial and temporal dimensions. These three types of fusion strategies can also be generalized to other learning-based action representation approaches. Details of such a taxonomy will be given in the discussion subsection below. Due to the high computational requirements of CNN-based approaches, especially when being extended to incorporate the additional dimension, efforts have also been paid on how to speedup the training process in [92], where a modified architecture that adopts both the context stream and the fovea stream is used in the implementation. Specially, in the context stream, the input video clips are down sampled to half of the original spatial resolution,

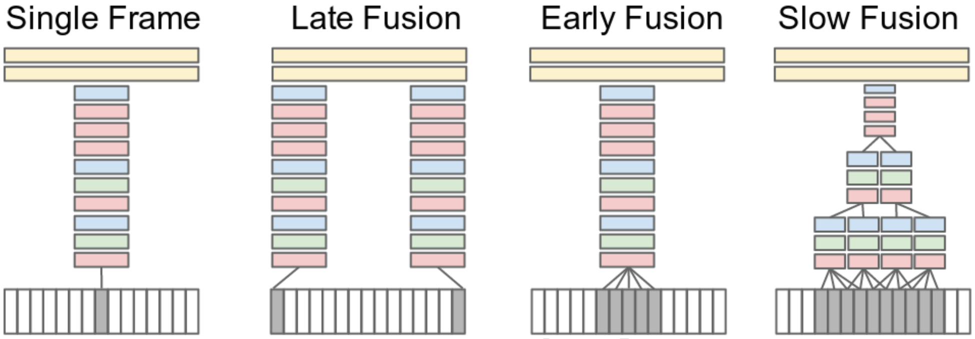

Figure 4: Illustration of different fusion strategies for incorporating with the temporal dimension (reprint from Karpathy *et al.* [92]).

while in the fovea stream, the centering regions of the original frame are cropped at an identical size of the down sampled resolution as in the context stream. The activations from both streams are concatenated and fed into the first fully connected layer for generating the final action representation. Interestingly, experimental results suggest that performing CNNs on individual static video frames achieves similar performance as performing CNNs on a stack of frames, which suggest that motion information is not well captured by the proposed approach.

3.4. Hybrid models

Different from the two-stream architecture of the multiresultion strategy in [92] that aims to speed up the training process with such a architecture design, a two-stream convolutional networks approach is recently proposed by Simonyan and Zisserman [93] to learn action representations by decomposing action videos into both spatial and temporal components. In stead of implicitly estimating the motion as in [92] and [58], the temporal stream architecture takes a stack of optical flow displacement fields [94] between several adjacent frames as inputs, and train the convolutional networks on the top of optical flows. Several configurations of convolutional networks are further proposed in [93], including the optical flow stacking, which uses dense optical flows of neighboring pairs of consective frames, the trajectory stacking, which samples the flow across several frames along the motion trajectories, and the bi-directional optical flow, which computes both the forward and backward directions of optical flows in either a dense manner or a trajectory-based sampling manner. Training the convolutional networks is conducted in a multi-task learning setting [95], where the task is set to classify videos in both the HMDB-51 dataset [96] and the UCF-101 dataset [97] by equipping two softmax classifcation layers on the top of the last fully connected layer, so that one layer computes the HMDB-51 classification score and the other computes the UCF-101 classification score. The training procedures of the both streams are generally the same, and similar to the network architectures of [98] and [99], where the temporal network can be considered as an adaptation of [100] to the temporal domain. While the spatial network performs 2D convolutions on static action frames, the 3D convolution architecture is also similar to the architectures used in [56], [58] and [92]. In terms of how convolutional networks are applied to the video data, the proposed two-stream approach can be considered as a hybrid model of learning from both handcrafted features and raw pixels.

Motivated by [93] and the success of the shallow handcrafted video representation [45], Wang *et al.* [101] propose a trajectory-pooled two-stream deep convolution descriptor (TDD). By adopting a similar network architecture as the two-stream learning architecture in [93] and performing training on the combination of the HMDB-51 dataset [96] and the UCF-101 dataset, the trained convolutional networks can be considered as generic feature extractors, which are then used to compute multi-scale convolutional feature maps from each video in the target dataset. Meanwhile, improved trajectories are computed using the method in [45] from each action video. Trajectory-pooling is then performed by pooling local convolutional network responses along the trajectories in the spatio-temporal space, where the final TDD descriptors are obtained from the pooling results. Compared to [93], [45] inclines more towards the handcrafted feature side, since that in addition to the optical flows used when training the two-stream network, dense trajectories are utilized for pooling the deep convolutional descriptors. On the top of TDDs, Wang *et al.* [101] further leverage the discriminative Fisher vector [102] for encoding the TDDs, where their reported leading performance is achieved by early fusing TDD and improved dense trajectory features using the Fisher vector representations. However, since fisher vector representation is not used for obtaining the classification results in [93], there is no evidence to support the superiority in terms of network architectures of TDD over the original two-stream convolutional networks [93].

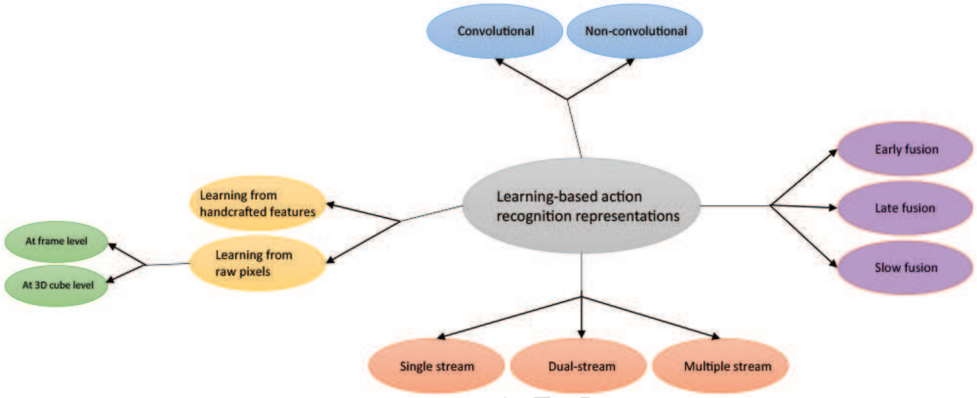

Figure 5: Expanded taxonomies of learning-based action recognition representations.

3.5. Discussion

The above mentioned learning-based action representations cover the majority of most popular approaches that are based on network architectures and published after 2010. Based on how learnings are performed over action videos, we put these approaches in multiple branches, including learning from static frames, learning from frame transformations, learning from handcrafted features, learning with three dimensional neural networks and learning with hybrid models. In fact, these taxonomies overlap with each other. For example, the works [93] and [101] are categorized as hybrid models, while their learning architectures also take handcrafted features as inputs; the unsupervised learning approach in [72] also cooperates with the three-dimensional convolutional network; and in addition to [72], the majority of above mentioned approaches include elements that aim to learn from action motions. Thus, some other taxonomies can be used for discriminating different learning-based action representations. We illustrate these expanded taxonomies in Figure 5, and list the corresponding belongingness of learning-based action representations.

The majority of popular existing learning-based action representations are either directly applying CNN to video frames or variants of CNN for learning spatio-temporal features. Network architectures of some popular learning-based representations that are established based on CNN are listed in Table 3. For other learning-based methods [80], [57], since the CNN architecture is not utilized in their learning frameworks, we do not list them in Table 3. The listed network architectures vary in the input data types,

Table 2: Performance comparisons between state-of-the-art learning-based action representations and handcrafted representations

Methods	Dataset	Performance
STF Taylor *et al.* [72]	KTH [104]	90.0%
SS (Short Sequence) Schindler and Gool [103]		95.04%
BIS Jhuang *et al.* [80]		91.70%
SDL Baccouche *et al.* [58]		94.39%
STIP+BoW Kuehne *et al.* [96]	UCF-101 [97]	43.9%
Motionlets Wang *et al.* [105]		63.3%
DT+BoW Wang *et al.* [106]		79.9%
DT+MVSV Cai *et al.* [107]		83.5%
IDT+FV Wang *et al.* [45]		85.9%
IDT+HSV Peng *et al.* [108]		87.9%
SFCNN Karpathy *et al.* [92]		65.4%
Two-stream+SVM Simonyan and Zisserman [93]		88.0%
TDD+FV Wang *et al.* [101]		90.3%
TDD+IDT+FV Wang *et al.* [101]		91.5%
Feature pruning Liu *et al.* [109]	YouTube	71.2%
HIST Le *et al.*, [57]		75.8%
IDT+SFV Peng *et al.* [51]		93.38%
HOG/HOF + KM + SVM Wang *et al.* [40]	Hollywood2	47.4%
VideoDarwin Fernando *et al.* [110]		73.7%
IDT+FV Wang *et al.* [45]		64.3%
STF Taylor *et al.* [72]		46.6%
STIP+BoW Kuehne *et al.* [96]	HMDB-51 [96]	23.0%
Motionlets Wang *et al.* [105]		42.1%
DT+BoW Wang *et al.* [106]		46.6%
DT+MVSV Cai *et al.* [107]		55.9%
IDT+FV Wang *et al.* [45]		57.2%
IDT+HSV Peng *et al.* [108]		61.1%
Two-stream+SVM Simonyan and Zisserman [93]		63.2%
TDD+FV Wang *et al.* [101]		63.2%
TDD+IDT+FV Wang *et al.* [101]		65.9%
VideoDarwin Fernando *et al.* [110]		63.7%
IDT+SFV Peng *et al.* [51]		66.79%

layer numbers, layer components and orders of how different components are connected. So far, there does not exist a widely acknowledged network architecture for learning action representations. The SFCNN approach utilizes the maximum layer numbers among listed works, however, the proposed network achieves equivalent results as the network that operates on individual video frames, which suggest that motion information is not effectively learned in such a network. Also, based on the fact that the SFCNN achieves a significantly worse result than the best handcrafted shallow representations [45],[51], SFCNN fails to demonstrate that a network with long cascaded layers can lead to better performance. The network architectures used in both [93] and [101] are based on the image visualizing and understanding network [99], and are perceptually effective since they achieve leading performance among learning-based action representations. Unfortunately, since the motion information is still captured by handcrafted features in both approaches, and the reported result on the spatial stream network is significantly worse than either the best handcrafted shallow representations or their proposed two-stream model, there is no proof that the improved performance in [93] and [101] come from the superiority of the network architecture.

Unlike spatial CNNs, which can be trained using large-scale image data (such as ImageNet [17]), most available datasets for action classification are still rather small, e.g., the UCF-101 dataset [97], which consists of 9.5K videos from 101 categories, and the HMDB-51 dataset [96], which consists of 3.7K videos from 51 categories. In order to increase the number of available videos for training, a straightforward option is to simply combine both datasets for training, however, this is not practical due to the intersection between the sets of classes. Consequently, recent approaches [93], [92] seek more principled ways of combining existing action datasets using either multi-task learning [95],[111],[112] or transfer learning [113],[114],[115]. In the multi-task learning setting of [95], the architecture of the convolutional network is modified so that 2 softmax classification layers can be cascaded to the top of the layer fully connected layer, where one soft layer handles the classification task of the HMDB-51 dataset while the other handles the classification task in the UCF-101 dataset. On the other hand, in the transfer learning setting employed by Wang *et al.* [101], the model is trained on the UCF-101 dataset, and extended to the HMDB-51 dataset for action feature extraction and action classification. So far, the largest available video dataset is the Sports-1M dataset [92], which consists of 1 million YouTube videos annotated with 487 categories. As claimed in [92], since the YouTube video IDs of UCF-101

Table 3: Inputs and network architectures of some popular action feature learning approaches that are established on CNN. Layer numbers are reported by including the input layers for all methods. Definitions of notations used for describing network architectures are as follows: "H" denotes the hardwired layer, "C" denotes the convolutional layer, "N" denotes the normalization layer, "S" denotes the sub-sampling layer, "R" denotes the rectification layer, "P" denotes the pooling layer and "F" denotes the fully connected layer.

Method	Input	Layer number	Architecture
3D CNN Ji *et al.* [56]	Action videos	7	H⇒C⇒S⇒C⇒S⇒C⇒F
MCNN Kim *et al.* [77]	Action descriptors obtained from 3D Gabor filter	5	H⇒C⇒S⇒C⇒S⇒
SDL Baccouche *et al.* [58]	Action videos	10	H⇒C×2⇒R×2⇒S×2⇒C⇒R⇒S⇒C⇒N⇒N
SFCNN Karpathy *et al.* [92]	Multiresolution Action videos	13	H⇒C⇒N⇒P⇒C⇒N⇒P⇒C⇒C⇒C⇒P⇒F⇒F
Two-stream Simonyan and Zisserman [93]	Action videos and optical flows	8	H⇒C⇒C⇒C⇒C⇒C⇒F⇒F
TDD Wang *et al.* [101]	Trajectory-pooled action videos and optical flows	8	H⇒C⇒C⇒C⇒C⇒C⇒F⇒F

videos are not available, there is no rigorous guarantee that the HMDB-51 dataset and UCF-101 dataset do not overlap with each other. Unfortunately, even by ruling out the dataset intersections between the Sports-1M dataset and the UCF-101 dataset, the model trained on the Sports-1M dataset is still not able to lead the best performance.

4. Conclusion

In this survey, we provide a comprehensive study over the state-of-the-art action representations, covering both handcrafted representations and learning-based representations. Due to the success of deep learning in image categorization tasks, increasing attentions have been paid on migrating the deep learning architectures from image-level visual categorization tasks to video-level action sequences recognition tasks. On the handcrafted representation side, we deliver both the foundations of classical action representations and recent advancements in action representations. So far, handcrafted action features have achieved remarkable performance on a variety of action recognition tasks and made significant contributions to the action recognition society. However, since the most effective handcrafted action features follow a densely-sampled local strategy, high computational complexity is unavoidable in both training and testing phases, which make these handcrafted approaches inapplicable for real-world applications. Learning-based approaches, on the other hand, can avoid such high computation by employing simple network architectures with learned parameters. With the rapid developments of learning-based action representations, the most effective learning-based approach can outperform the best performance achieved by handcrafted action features, though these learning-based approaches still rely on handcrafted features. However, the performance of pure learning-based approaches that directly from videos still fall far behind action recognition researchers' expectations. The reasons can be two folds: 1) unlike those successful cases in the image domain, the majority of existing learning-based approaches are not trained on large-scale data, which can result in the insufficiency of the networks' generalization ability; 2) in order to balance the computational costs, down-sampling strategies are widely utilized in learning-based action representations, where the information loss during the down-sampling procedures could cause different levels of performance degradation. In the future, we can anticipate the emergence of more advanced action representations that are able to achieve improved performance while remaining the simplicity in terms of network architectures.

5. Reference

[1] D. G. Daniel Schacter, D. Wegner, Psychology, New York: Worth, 2011.

[2] G. Buccino, F. Binkofski, L. Riggio, The mirror neuron system and action recognition, Brain and language 89 (2) (2004) 370–376.

[3] F. Pulvermüller, Y. Shtyrov, R. Ilmoniemi, Brain signatures of meaning access in action word recognition, Journal of cognitive neuroscience 17 (6) (2005) 884–892.

[4] C. Keysers, E. Kohler, M. A. Umiltà, L. Nanetti, L. Fogassi, V. Gallese, Audiovisual mirror neurons and action recognition, Experimental brain research 153 (4) (2003) 628–636.

[5] T. Spencer, F. Bazer, Biology of progesterone action during pregnancy recognition and maintenance of pregnancy., Frontiers in bioscience: a journal and virtual library 7 (2002) d1879–98.

[6] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[7] A. Klaser, M. Marszałek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: British Machine Vision Conference, British Machine Vision Association, 2008.

[8] M. D. Rodriguez, J. Ahmed, M. Shah, Action mach a spatio-temporal maximum average correlation height filter for action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[9] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, R. Moore, Real-time human pose recognition in parts from single depth images, Communications of the ACM 56 (1) (2013) 116–124.

[10] J. K. Aggarwal, M. S. Ryoo, Human activity analysis: A review, ACM Computing Surveys (CSUR) 43 (3) (2011) 16.

[11] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, B. P. Buckles, Advances in Human Action Recognition: A Survey, arXiv preprint arXiv:1501.05964 .

[12] T. B. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, Computer vision and image understanding 104 (2) (2006) 90–126.

[13] R. Poppe, A survey on vision-based human action recognition, Image and vision computing 28 (6) (2010) 976–990.

[14] J. K. Aggarwal, Q. Cai, Human motion analysis: A review, Computer vision and image understanding 73 (3) (1999) 428–440.

[15] C. Cedras, M. Shah, Motion-based recognition a survey, Image and Vision Computing 13 (2) (1995) 129–155.

[16] D. M. Gavrila, The visual analysis of human movement: A survey, Computer vision and image understanding 73 (1) (1999) 82–98.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition,, 2009.

[18] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (3) (2011) 27.

[19] D. Weinland, M. Özuysal, P. Fua, Making action recognition robust to occlusions and viewpoint changes, in: European Conference on Computer Vision, Springer, 2010.

[20] V. A. Diwadkar, T. P. McNamara, Viewpoint dependence in scene recognition, Psychological Science 8 (4) (1997) 302–307.

[21] D. Weinland, R. Ronfard, E. Boyer, Free viewpoint action recognition using motion history volumes, Computer Vision and Image Understanding 104 (2) (2006) 249–257.

[22] R. Souvenir, J. Babbs, Learning the viewpoint manifold for action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition,, 2008.

[23] D. Weinland, E. Boyer, R. Ronfard, Action recognition from arbitrary views using 3d exemplars, in: IEEE International Conference on Computer Vision,, 2007.

[24] V. Parameswaran, R. Chellappa, View invariance for human action recognition, International Journal of Computer Vision 66 (1) (2006) 83–101.

[25] L. Xia, C.-C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3d joints, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2012.

[26] I. Laptev, On space-time interest points, International Journal of Computer Vision 64 (2-3) (2005) 107–123.

[27] G. Willems, T. Tuytelaars, L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: European Conference on Computer Vision, Springer, 2008.

[28] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: ACM International Conference on Multimedia, 2007.

[29] A. Gilbert, J. Illingworth, R. Bowden, Scale invariant action recognition using compound features mined from dense spatio-temporal corners, in: European Conference on Computer Vision, Springer, 2008.

[30] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.

[31] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance,, 2005.

[32] W. Förstner, E. Gülch, A fast operator for detection and precise location of distinct points, corners and centres of circular features, in: Intercommission Conference on Fast Processing of Photogrammetric Data, 1987.

[33] C. Harris, M. Stephens, A combined corner and edge detector., in: Alvey vision conference, vol. 15, Citeseer, 50, 1988.

[34] G. Sotak, K. L. Boyer, The Laplacian-of-Gaussian kernel: a formal analysis and design procedure for fast, accurate convolution and full-frame output, Computer Vision, Graphics, and Image Processing 48 (2) (1989) 147–189.

[35] R. Chaudhry, A. Ravichandran, G. Hager, R. Vidal, Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions, in: IEEE Conference on Computer Vision and Pattern Recognition,, 2009.

[36] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition,, 2005.

[37] J. Sivic, A. Zisserman, Video Google: A text retrieval approach to object matching in videos, in: IEEE International Conference on Computer Vision,, 2003.

[38] M. Aharon, M. Elad, A. Bruckstein, ¡ img src=”/images/tex/484. gif” alt=”\ rm K”¿-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation, IEEE Transactions on Signal Processing, 54 (11) (2006) 4311–4322.

[39] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[40] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: British Machine Vision Conference, 2009.

[41] V. Ganapathi, C. Plagemann, D. Koller, S. Thrun, Real time motion capture using a single time-of-flight camera, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[42] M. Siddiqui, G. Medioni, Human pose estimation from a single view point, real-time range sensor, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2010.

[43] Y. Zhu, K. Fujimura, Constrained optimization for human pose estimation from depth sequences, in: Asian Conference on Computer Vision, Springer, 2007.

[44] J. L. Wilson, Microsoft kinect for Xbox 360, PC Mag. Com .

[45] H. Wang, C. Schmid, Action recognition with improved trajectories, in: IEEE International Conference on Computer Vision, 2013.

[46] N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of flow and appearance, in: European Conference on Computer Vision, Springer, 2006.

[47] E. Vig, M. Dorr, D. Cox, Space-variant descriptor sampling for action recognition based on saliency and eye movements, in: European Conference on Computer Vision, Springer, 2012.

[48] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, C.-W. Ngo, Trajectory-based modeling of human actions with motion reference points, in: European Conference on Computer Vision, Springer, 2012.

[49] H. Wang, C. Schmid, Action recognition with improved trajectories, in: IEEE International Conference on Computer Vision, 2013.

[50] H. Bay, T. Tuytelaars, L. Van Gool, Surf: Speeded up robust features, in: European Conference on Computer vision, Springer, 2006.

[51] X. Peng, C. Zou, Y. Qiao, Q. Peng, Action recognition with stacked fisher vectors, in: European Conference on Computer Vision, Springer, 581–595, 2014.

[52] F. Perronnin, J. Sánchez, T. Mensink, Improving the fisher kernel for large-scale image classification, in: European Conference on Computer Vision, Springer, 2010.

[53] X. Zhen, L. Shao, X. Li, Action recognition by spatio-temporal oriented energies, Information Sciences 281 (2014) 295–309.

[54] L. Shao, X. Zhen, D. Tao, X. Li, Spatio-temporal Laplacian pyramid coding for action recognition, IEEE Transactions on Cybernetics 44 (6) (2014) 817–827.

[55] S. Marčelja, Mathematical description of the responses of simple cortical cells*, JOSA 70 (11) (1980) 1297–1300.

[56] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (1) (2013) 221–231.

[57] Q. V. Le, W. Y. Zou, S. Y. Yeung, A. Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[58] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, A. Baskurt, Sequential deep learning for human action recognition, in: Human Behavior Understanding, Springer, 29–39, 2011.

[59] W. Banzhaf, P. Nordin, R. E. Keller, F. D. Francone, Genetic programming: an introduction, vol. 1, Morgan Kaufmann San Francisco, 1998.

[60] L. Shao, L. Liu, X. Li, Feature learning for image classification via multiobjective genetic programming, IEEE Transactions on Neural Networks and Learning Systems 25 (7) (2014) 1359–1371.

[61] C. S. Burrus, R. A. Gopinath, H. Guo, Introduction to wavelets and wavelet transforms: a primer .

[62] Z. S. Harris, Distributional structure, Word 10 (2-3) (1954) 146–162.

[63] S. Jones, L. Shao, Unsupervised spectral dual assignment clustering of human actions in context, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 604–611, 2014.

[64] T. Guha, R. K. Ward, Learning sparse representations for human action recognition, Pattern Analysis and Machine Intelligence, IEEE Transactions on 34 (8) (2012) 1576–1588.

[65] J. Zheng, Z. Jiang, P. J. Phillips, R. Chellappa, Cross-View Action Recognition via a Transferable Dictionary Pair., in: British Machine Vision Conference, 2012.

[66] F. Zhu, L. Shao, Weakly-supervised cross-domain dictionary learning for visual recognition, International Journal of Computer Vision 109 (1-2) (2014) 42–59.

[67] Y. Bengio, Y. LeCun, et al., Scaling learning algorithms towards AI, Large-scale kernel machines 34 (5).

[68] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks, arXiv preprint arXiv:1312.6229 .

[69] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional Architecture for Fast Feature Embedding, arXiv preprint arXiv:1408.5093 .

[70] F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou, P. E. Barbano, Toward automatic phenotyping of developing embryos from videos, IEEE Transactions on Image Processing 14 (9) (2005) 1360–1371.

[71] G. E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural computation 18 (7) (2006) 1527–1554.

[72] G. W. Taylor, R. Fergus, Y. LeCun, C. Bregler, Convolutional learning of spatio-temporal features, in: European Conference on Computer Vision, Springer, 2010.

[73] R. Memisevic, G. Hinton, Unsupervised learning of image transformations, in: IEEE Conference on Computer Vision and Pattern Recognition,, 2007.

[74] C. Cadieu, B. A. Olshausen, Learning transformational invariants from natural movies, in: Advances in neural information processing systems, 2008.

[75] Y. LeCun, K. Kavukcuoglu, C. Farabet, Convolutional networks and applications in vision, in: IEEE International Symposium on Circuits and Systems, 2010.

[76] K. Jarrett, K. Kavukcuoglu, M. Ranzato, Y. LeCun, What is the best multi-stage architecture for object recognition?, in: IEEE International Conference on Computer Vision, 2009.

[77] H.-J. Kim, J. S. Lee, H.-S. Yang, Human action recognition using a modified convolutional neural network, in: Advances in neural information processing systems, Springer, 2007.

[78] J. P. Jones, L. A. Palmer, An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex, Journal of neurophysiology 58 (6) (1987) 1233–1258.

[79] H.-J. Kim, J. Lee, H.-S. Yang, A weighted FMM neural network and its application to face detection, in: Neural Information Processing, Springer, 177–186, 2006.

[80] H. Jhuang, T. Serre, L. Wolf, T. Poggio, A biologically inspired system for action recognition, in: IEEE International Conference on Computer Vision,, 2007.

[81] K. Fukushima, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biological cybernetics 36 (4) (1980) 193–202.

[82] J. Mutch, D. Lowe, Multiclass object recognition using sparse, localized hmax features, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006.

[83] T. Serre, L. Wolf, T. Poggio, Object recognition with features inspired by visual cortex, in: IEEE Conference on Computer Vision and Pattern Recognition,, 2005.

[84] D. Wu, L. Shao, Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[85] G. W. Taylor, G. E. Hinton, S. T. Roweis, Modeling human motion using binary latent variables, in: Advances in neural information processing systems, 2006.

[86] L. E. Baum, T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, The annals of mathematical statistics .

[87] A. Hyvärinen, J. Hurri, P. O. Hoyer, Natural Image Statistics: A Probabilistic Approach to Early Computational Vision., vol. 39, Springer Science & Business Media, 2009.

[88] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics and intelligent laboratory systems 2 (1) (1987) 37–52.

[89] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, J. Schmidhuber, A novel connectionist system for unconstrained handwriting recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (5) (2009) 855–868.

[90] F. A. Gers, N. N. Schraudolph, J. Schmidhuber, Learning precise timing with LSTM recurrent networks, The Journal of Machine Learning Research 3 (2003) 115–143.

[91] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1110–1118, 2015.

[92] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[93] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014.

[94] T. Brox, A. Bruhn, N. Papenberg, J. Weickert, High accuracy optical flow estimation based on a theory for warping, in: European Conference on Computer Vision, Springer, 2004.

[95] R. Collobert, J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in: ACM International Conference on Machine Learning, 2008.

[96] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: a large video database for human motion recognition, in: IEEE International Conference on Computer Vision, 2011.

[97] K. Soomro, A. R. Zamir, M. Shah, Ucf101: A dataset of 101 human actions classes from videos in the wild, arXiv preprint arXiv:1212.0402 .

[98] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, arXiv preprint arXiv:1405.3531 .

[99] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer, 2014.

[100] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012.

[101] L. Wang, Y. Qiao, X. Tang, Action Recognition With Trajectory-Pooled Deep-Convolutional Descriptors, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[102] J. Sánchez, F. Perronnin, T. Mensink, J. Verbeek, Image classification with the fisher vector: Theory and practice, International Journal of Computer Vision 105 (3) (2013) 222–245.

[103] K. Schindler, L. Van Gool, Action snippets: How many frames does human action recognition require?, in: IEEE Conference on Computer Vision and Pattern Recognition,, 2008.

[104] C. Schüldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: IEEE International Conference on Pattern Recognition,, 2004.

[105] L. Wang, Y. Qiao, X. Tang, Motionlets: Mid-level 3d parts for human motion recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[106] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Dense trajectories and motion boundary descriptors for action recognition, International journal of computer vision 103 (1) (2013) 60–79.

[107] Z. Cai, L. Wang, X. Peng, Y. Qiao, Multi-view super vector for action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[108] X. Peng, L. Wang, X. Wang, Y. Qiao, Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice, arXiv preprint arXiv:1405.4506 .

[109] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition,, 2009.

[110] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, T. Tuytelaars, Modeling video evolution for action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[111] Y. Yan, G. Liu, E. Ricci, N. Sebe, Multi-task linear discriminant analysis for multi-view action recognition, in: IEEE International Conference on Image Processing, 2013.

[112] Q. Yang, Activity Recognition: Linking Low-level Sensors to High-level Intelligence., in: International Joint Conferences on Artificial Intelligence, 2009.

[113] V. W. Zheng, D. H. Hu, Q. Yang, Cross-domain activity recognition, in: ACM International Conference on Ubiquitous Computing, 2009.

[114] J. Liu, M. Shah, B. Kuipers, S. Savarese, Cross-view action recognition via view knowledge transfer, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[115] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.

Highlights

1. We provide a comprehensive stud over the state-of-the-art action representations.
2. Deep learning-based representations are compared to handcrafted representations.
3. Pros and cons of current deep learning-based approaches are discussed.